25
22
20
28
32
36
40
45
50
55
100 mm
1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
4.5
5.0
5.6
2.5
2.2
2.0
1.8
ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz1234567890
ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz
1234567890
1.0 mm
1.5 mm
2.0 mm
A5

DOCUMENT RESUME

ED 290 795 TM 870 586

AUTHOR Littlefield, John H.; Troendle, G. Roger
TITLE Effects of Rating Task Instructions on Consistency and Accuracy of Expert Raters.
PUB DATE Apr 87
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Dentistry; Evaluation Methods; Higher Education; *Holistic Evaluation; *Interrater Reliability; *Scoring; *Test Reliability
IDENTIFIERS *Rater Reliability

ABSTRACT

The effect of different types of rating task instructions on rater behavior was examined using experts, as opposed to novices, as raters. The experts were instructed to (1) form a global categorical judgment (early hypothesis generation); (2) assess 19 detailed elements; or (3) both. Subjects were 8 dental faculty members who ranged in age from 28 to 60 and who had at least 2 years of teaching experience. The task was to evaluate five three-fourth dental crown preparations twice, using each of the three types of rating instructions. Intrarater and interrater agreement and reliability were assessed, as was the level of rater accuracy. Higher coefficients of rater reliability, but not agreement, resulted from the global and combined instructions, compared with the 19-point, or traditional, instruction. The global judgment alone improved reliability over traditional instructions, but intrarater agreement was lower. Expert rater consistency was higher when early hypothesis generation and self-monitoring were encouraged by the rating instructions. There were no significant differences in score accuracy among the three types of instruction. (MGD)

Effects of Rating Task Instructions on Consistency and Accuracy of Expert Raters

John H. Littlefield, Ph.D. and G. Roger Troendle, D.D.S., M.S.

University of Texas Health Science Center at San Antonio

Paper Presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., April, 1987

Effects of Rating Task Instructions on Consistency and Accuracy of Expert Raters

Background

Following a comprehensive review of the performance rating research, Landy and Farr (1980) recommended a moratorium on rating form research. They noted that the rater's information processing serves as a cognitive filter of the measurement data and that we need to better understand the cognitive processes raters use in making judgments. Almost synchronously, major theoretical articles appeared addressing issues such as cognitive models to account for halo error (Cooper, 1981) and automatic vs. consciously-controlled rater judgment (Feldman, 1981). Sophisticated cognitive research is now being conducted to improve our understanding of the influence of relevant rater knowledge on halo error (Kozlowski et. al., 1986) and rater cognitive simplification strategies (Cadwell and Jenkins, 1986). Cadwell and Jenkins describe the "rater as the measuring instrument," reflecting a profound shift in the focus of performance ratings research. This study will use cognitive processing of experts as a conceptual framework to predict changes in rater behavior in response to different types of rating task instructions.

Research in cognitive psychology has expanded our understanding of cognitive processing by experts. Unlike novices, experts form highly elaborated cognitive representations of a problem (Larkin et. al., 1980). Experts organize knowledge structures over long periods of learning and experience (Glaser, 1984). When faced with a problem, experts automatically (i.e., without conscious effort) construct an initial high quality representation of the problem. Their knowledge is "chunked" around principles and abstractions which subsume surface features of the problem and their perceptions are influenced by pattern recognition processes (Brandsford, et. al., 1986). In medical diagnosis, for example, expert physicians generate hypotheses early-on, and the correct diagnosis is very likely among those hypotheses (Norman, 1985). In the public schools, teachers "size up" students as individuals, grouping them very quickly, and these initial estimates remain quite stable (Stiggins, et. al, 1986). Cognitive psychology research is providing an insight into the power of human thinking to use a large knowledge base in an efficient and automatic manner.

Most performance ratings research has used nonexpert raters as subjects, typically college students, in real life training environments, however, raters are frequently experts at the tasks being rated (e.g., physicians, teachers). When judging a performance, an expert rater is likely to quickly construct an initial representation of the performance. That representation will include knowledge about the appropriateness of various performance elements to solving the problem at hand. For example, in conducting a physical exam, a medical student may follow the correct procedure but overlook a disease finding, thereby seriously compromising the validity of the entire exam. An expert rater would judge the performance as a failure while a novice rater would simply note a step that was overlooked.

Recognizing that raters are experts suggests a very different role for the performance rating form. Instead of being an "instrument" which defines how observations are to be made, the rating form could be designed to facilitate communicating and quantifying observations by the expert. The instructions on the proposed rating form would request a global categorical judgment (early hypothesis generation) plus assessment of various detailed elements of the performance. The detailed assessment would serve as a stimulus for rater self-monitoring to verify the initial "early hypothesis" and also provide documentation for the rationale used in making the global judgment.

The general appearance of the rating form proposed above is not a radical departure from traditional formats; its role in relation to rater cognitive processes is significantly changed however. Traditional instructions to the rater are to observe the performance and mark numerous detailed criteria in a mechanical fashion. Marking the detailed criteria parallels a novice's approach to problem solving by collecting numerous miscellaneous facts (Larkin et. al., 1980). Traditional rating instructions include calculating a score by summing the marks. This scoring approach reflects psychometric test theory, which attributes considerable specificity and measurement error to

individual items, the summed score therefore will be a more generalizable measure of overall performance (Nunally, 1978). The proposed instructions emphasize the expert's global judgment with the assessment of detailed performance elements serving an ancillary role.

If the proposed rating task instructions facilitate quantifying expert observations, the resultant ratings should display improved consistency and accuracy when compared to traditional rating instructions. Rater consistency has two components: (1) agreement - the extent to which different judges make exactly the same judgment; and (2) reliability - the degree to which ratings by different judges are proportional when expressed as deviations from their respective mean scores (Tinsley and Weiss, 1975). Hogan et. al. (1986) have recently demonstrated the importance of calculating both measures of rater consistency in an applied setting. Lord (1983) recommends that rater accuracy is best assessed by calculating Receiver Operating Curves (ROC). ROC analysis graphically displays the trade-off between the probability of making true positive vs. false positive decisions. In applied decision-making settings, one must always judge between the relative cost of making false positive vs. false negative errors.

Methods

Subjects were eight dental faculty members who ranged in age from 28 to 60 years. All subjects had two or more years of clinical teaching experience and had participated in developing the detailed rating criteria used in this study.

The rating task in this experiment is a routine part of the subjects' daily job responsibilities. The stated purpose of the experimental task was to standardize grading methods. The task was to evaluate five 3/4 crown preparations twice using each of three different types of rating instructions.

1. *Traditional* -mark each of 19 criteria on a three-category scale (Acceptable, needs Improvement or Unsatisfactory). "Be sure to mark each criterion either A, I, or U." A single composite score was calculated ex post facto by the investigators.
2. *Global Judgment Only* - no detailed criteria were available. "After inspecting the tooth, write your grade (4,3,2,1,0). As in clinic, 4 is the best grade and 0 is a failure."
3. *Combined* -"After marking the criteria, assign a grade according to the grade code provided."

The *Combined* instructions condition is an attempt to conform the rating task instructions to the cognitive processes of experts. Presumably, the rater would initially form an early hypothesis about the tooth being judged then review the detailed criteria to confirm the hypothesis. The *Global Judgment Only* condition provides a middle ground between *Traditional* rater instructions and the *Combined* instructions. *Global Judgment* should improve rater consistency and accuracy over the *Traditional* instructions, but the addition of detailed criteria in the *Combined* condition should improve consistency and accuracy even further by serving as a self-monitoring check for the expert raters to verify their initial hypotheses.

Data collection procedures were described in detail by Troendle (1983). Raters were assigned code numbers to maintain anonymity and were not informed that they were reevaluating the same five teeth. Teeth were identified only by code numbers, and at least six weeks intervened between each trial session.

Data analysis procedures addressed two general questions:

1. Do the three types of rating instructions result in different levels of *intra* rater and *inter* rater consistency?
2. Do the three types of rating instructions result in different levels of rater accuracy?

As noted earlier, rater consistency has two components: agreement and reliability. Intrarater and interrater agreement were assessed using a *tau* coefficient suggested by Tinsley and Weiss (1975). *Tau* consists of a Chi Square test to ensure that observed agreement exceeds chance levels followed by calculation of percent agreement coefficient adjusted down for chance agreements. Statistical significance of differences in agreement levels among the three types of rating instructions were

tested by calculating Cochran's Q (SPSS, 1984) using actual (unadjusted) frequency of agreement data. Intrarater reliability was assessed using Kendell's tau-b correlation coefficient while interrater reliability was assessed using an intraclass correlation coefficient recommended by Finn (1970). Rater accuracy was assessed by calculating three Receiver Operating Curves (ROC) based upon Signal Detection Theory (Swets and Pickett, 1982). Data from the pairs of trials (1-2, 3-4, & 5-6) were pooled to calculate the ROC for each type of rating instructions. Statistical significance of differences in accuracy among the three types of rating instructions was assessed by calculating critical ratios among the various pairs of the three curves (Metz, et. al., 1984). Interrater consistency is the upper limit of rater accuracy in the same sense that test reliability is the upper bound of test validity.

Results

Figure 1 displays the design of the study, descriptive statistics and various intrarater and interrater consistency coefficients.

Figure 1 - Study Design, Descriptive Statistics and Consistency Coefficients

Instructions	Traditional		Global Only		Combined	
Trial Number	#1	#2	#3	#4	#5	#6
# of Raters	8	8	8	8	8	8
# of Teeth	5	5	5	5	5	5
Mean Rating	1.38	1.35	1.88	1.75	1.90	1.83
Std. Dev.	1.66	1.69	1.27	1.52	1.60	1.58
Agreement [1]						
a. Intrarater	50%		41%		63%	
b. Interrater	27%		29%		47%	
Reliability						
a. Intrarater	.50		.68		.81	
b. Interrater	.34		.54		.47	

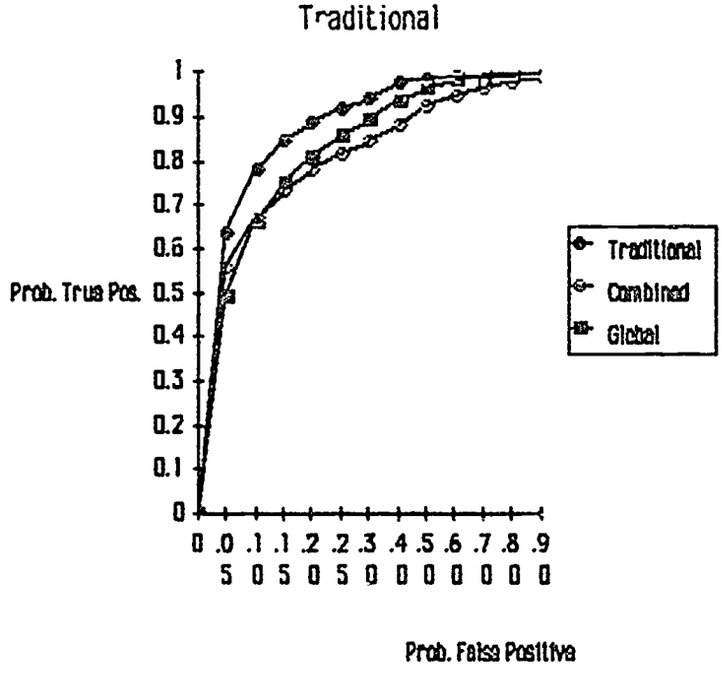
1. Adjusted down for agreements due to chance

Intrarater agreement was defined as rater *i* making the same judgment about tooth *j* both times the tooth was judged using a given set of rater instructions. Forty pairs of judgments were made under each condition. Levels of intrarater agreement among the three types of instructions did not differ significantly when the unadjusted frequencies of agreement were tested by Cochran's Q ($Q=3.84$, $p=.15$). Interrater agreement was defined as all 8 raters marking ± 1 category of the modal judgment for that tooth. A total of 10 teeth were judged under each condition. Levels of interrater agreement among the three types of instructions did not differ significantly when tested by Cochran's Q ($Q=1.14$, $p=.57$).

The intrarater reliability coefficients for the three types of rating instructions are substantially different. The upper limit of the 95% confidence interval for the *Traditional* instructions coefficient is .65 which does not include the other two coefficients. The tau-b coefficients were each based upon 40 matched pairs of judgments. The interrater reliability coefficients also are substantially different. The upper limit of the 95% confidence interval for the *Traditional* instructions coefficient is .52 which does not include the *Global Judgment Only* instructions coefficient. The intraclass correlation coefficients were calculated using the residual mean square from a one way ANOVA for each of the three groups of 80 scores (5 teeth x 8 raters x 2 trials).

Figure 2 presents Receiver Operating Curves (ROC) for each of the three types of rating instructions. Each curve is based upon 80 scores. To read the curves, note that at .1 probability of a false positive decision, the *Traditional* instructions result in a .78 probability of a true positive decision while the *Global Only* and the *Combined* instructions result in a .66 probability. One measure of accuracy using ROC analysis is the proportion of the unit square area covered by the curve (100% is perfect accuracy). The *Traditional* instructions curve covers 93 % of the area while the *Global Only* curve covers 89% and the *Combined* curve covers 87%. A correlated observations critical ratio test of the various pairs of curves (Metz, et. al., 1984) did not reject the null hypothesis that the various pairwise sets of rating data were samples drawn from the same underlying ROC curve. A visual inspection of the curves confirms the statistical finding.

Figure 2 - Receiver Operating Curves



Discussion

The *Combined* and *Global Judgment Only* instructions resulted in higher coefficients of rater reliability, but no difference in rater agreement when compared to the *Traditional* instructions. The higher reliability coefficients indicate greater discrimination power (i.e., more confidence in rank ordering the five teeth from best to worst). Mathematically, the differences are due to greater variance among raters judging the same tooth using *Traditional* instructions (i.e., the within tooth mean square is larger). The differences between intrarater reliability of *Combined* and *Traditional* instructions also exist when scoring is done by summing judgments of the detailed criteria (Littlefield and Troendle, 1986) thus the differences in rater reliability apparently are not an artifact of the categorical scoring method. Taken together, the higher rater reliability coefficients indicate that *Combined* and *Global Judgment Only* instructions cause expert raters to produce scores which are more numerically precise than *Traditional* instructions. Perhaps being instructed to "mark each criterion" with no reference to an overall judgment (i.e., the *Traditional* instructions) disrupts the early hypothesis generation process which experts typically use in making judgments. Marking detailed criteria in the *Combined* instructions improved the intrarater reliability coefficient in comparison to *Global Judgment Only* (.81 vs. .68), but resulted in slightly lower interrater reliability (.47 vs. .54).

Levels of rater agreement were not significantly different among the three types of rating instructions. The intrarater agreement levels under *Combined* instructions were just short of statistical significance when tested against the *Global Judgment Only* instructions

(.63 vs. .41, p<.06) again reflecting a possible benefit due to rater self-monitoring when marking the detailed criteria. The lack of statistically significant differences among interrater agreement levels may be due to low statistical power (n=10 in each group). In general, the consistency coefficients in Figure 1 indicate that the *Combined* and *Global Judgment Only* instructions have moderate-to-high reliability, but moderate-to-low egreement. Hogan et. al. (1986) also found differences between reliability and agreement in ratings of nursing home patient disability and recommended calculating both indices. Future performance ratings research should assess both of these espects of rater consistency.

There were no statistically significant differences in rater accuracy, however, the statistical power of the tests was low. A power analysis of the data (Metz et. al., 1984) indicated only a .23 probability of detecting a signficant difference between the *Traditional* and *Combined* instructions curves. In order to achieve a .75 probability of detecting a statistically significant difference, approximately 300 judgements would be needed as compared to 80 in this study. Three of the teeth received a majority judgment of "clinically acceptable" (72-94% agreement) while two were judged to be "clinically unacceptable" (73 & 86% agreement). Future studies of rater accuracy should use a larger number of items to be judged with a wider diversity in judgment difficulty in order to increase the statistical power.

Ratings from the *Combined* and *Global Judgment Only* conditions correlate .74 with each other but only .59 and .56 respectively with *Traditional* ratings. The *Traditional* instructions result in more stringent decisions than the *Combined* and *Global Judgment Only* instructions (50% failure rate vs. 29% and 19%). Taken together, these correlations and differences in stringency of ratings from each condition support the internal validity of the study, namely that rather modest changes in rating task instructions affect the judgments of expert raters.

The conclusions from this study are marred by at least two weaknesses. First, the raters knew the data were for research purposes. Landy and Farr (1980) noted that ratings for administrative purposes will be more lenient than those for research purposes. Raters in this study may have performed differently if the scores were to be used to determine student grades. A second weakness is the failure to use a randomized block design. One could argue that the raters *learned* the teeth in rating them six times. The teeth were numbered with different-colored ink and tape for each trial and stored loosely in a box. Posthoc conversations with the raters did not indicate that they recognized the teeth. With a small cohesive group of subjects, attempting to have different groups simultaneously using different rating instructions was deemed unfeasible.

Future research in this area should use a larger number of stimuli with more diverse levels of judgment difficulty in order to improve the power of the tests of differences among the resulting Receiver Operating Curves. It might be advantageous to make the overall judgment in the *Combined* rating instructions independent of what is marked on the detailed criteria. One can never develop rating criteria which anticipate all possible outcomes; therefore, the printed criteria should be viewed as a sample of all possible criteria which could be related to the overall judgment. Construct psychology (Button, 1985) offers a possible technology for identifying the general constructs used by experts to make judgments in their field of expertise. With a better understanding of cognitive processing by expert raters, rating forms and their related instructions could be designed to more effectively facilitate quantifying and communicating judgments.

The results of this study address general classification decisions but do not address the problem of rater agreement in marking detailed criteria. Interrater agreement in marking the detailed criteria was 9% for the *Traditional* rating instructions, not significantly higher than agreement due to chance. For the *Combined* instructions, the agreement level was 36%. Neither level is very encouraging when viewed from the perspective of providing formative feedback to students to improve future performance.

Conclusions

This study suggests that giving expert raters instructions that request a global categorical judgment supplemented by marking detailed criteria results in higher intrarater and interrater reliability than instructions that focus on marking each detailed criterion without reference to the overall judgment (i.e., traditional instructions). Giving rating instructions that request only a global judgment improves reliability over traditional instructions, but intrarater agreement is somewhat lower than when both a global judgment and detailed criteria assessment are requested. The results are interpreted from a conceptual framework of early hypothesis generation and self-monitoring by experts. The pattern of consistency coefficients support a theoretical prediction of higher expert rater consistency when early hypothesis generation and self-monitoring are encouraged by the rating instructions (i.e., rating instructions which paralleled expert cognitive processing resulted in better reliability among expert raters). There were no significant differences in score accuracy among the three types of rating instructions.

Bibliography

Brandsford, J., Sherwood, R., Vye, N. & Rieser, J. (1986). Thinking and problem solving: research foundations. *Am. Psy.* 41(10): 1078-1089.

Button, E (1985). *Personal Construct Theory and Mental Health.* Brookline Books, Cambridge, MA., p. 3 - 56.

Cadwell, J. & Jenkins, J. (1986). Teachers' judgments about their students: the effect of cognitive simplification strategies on the rating process. *Am. Ed Res. J.*, 23(3): 460-475.

Cooper, W. (1981). Uniquitous halo. *Psy. Bull.*, 90(2): 218-244.

Feldman, J. (1981). Beyond attribution theory: cognitive processes in performance appraisal. *J. Applied Psy.*, 66: 127-148.

Finn, R.H. (1970). A note on estimating the reliability of categorical data. *Educ. & Psy. Meas.*, 30:71-76.

Glaser, R. (1984). Education and thinking: the role of knowledge. *Am. Psy.*, 39(2): 93-104.

Hogan, A., Smith, D., & Jameson, J. (1986). Functional assessment of nursing home patients. reliability and relevance. *Eval. & Hlth. Prof.*, 9(3): 339-360.

Kozlowski, S., Kirsch, M. & Chao, G. (1986). Job knowledge, ratee familiarity, conceptual similarity and halo error: an exploration. *J. App. Psy.*, 71(1): 45-49.

Landy, F.J. & Farr, J.L. (1980). Performance rating. *Psy. Bull.*, 87:72-107.

Larkin, J., McDermott, J., Simon, D. & Simon, H. (1980). Expert and novice performance in solving physics problems. *Science,* 208: 1335-1342.

Littlefield, J. & Troendle, R. (1986). Rating format effects on rater agreement and reliability. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April, 1986 (ERIC ED 271 483).

Lord, R. (1983). Accuracy in behavioral measurement: an alternative definition based on raters' cognitive schema and signal detection theory. *J. App. Psy.*, 70(1): 66-71.

Metz, C., Wang, H. & Kronman, H. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. In Deconinck, F. *Information Processing in Medical Imaging.* Martinus Nijhoff, Boston.

Norman, G. (1985). The role of knowledge in teaching and assessment of problem-solving. *J. Instructional Design,* 8(7)

Nunnally, J. C. (1978). *Psychometric Theory* , New York: McGraw-Hill, p. 84.

SPSS, (1984). SPSS/Pro: SPSS for the DEC Professional 350. McGraw-Hill Book Company, Chicago, Illinois, 60611

Stiggins, R., Conklin, N., & Bridgeford, N. (1986). Classroom assessment: a key to effective education. *Educ. Meas: Issues and Practice,* Summer, 1986: 5-17.

Swets, J. & Pickett, R. (1982). *Evaluation of Diagnostic Systems.* New York: Academic Press.

Tinsley, H. & Weiss, D. (1975). Interrater reliability and agreement of subjective judgments. *J. Coun. Psy.*, 22(4): 358-376.

Troendle, G. (1983). The effects of three rating forms on intrarater and interrater agreement and reliability in the rating of 3/4 crown preparations. Unpublished thesis, University of Texas Graduate School of Biomedical Sciences, San Antonio, Texas.